Mars: Near-Optimal Throughput with Shallow Buffers in Reconfigurable Datacenter Networks

Vamsi Addanki
Faculty of Electrical Engineering and Computer Science, TU Berlin, Germany

Chen Avin
School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel

Stefan Schmid
Faculty of Electrical Engineering and Computer Science, TU Berlin, Germany

ABSTRACT

The performance of large-scale computing systems often critically depends on high-performance communication networks. Dynamically reconfigurable topologies, e.g., based on optical circuit switches, are emerging as an innovative new technology to deal with the explosive growth of datacenter traffic. Specifically, *periodic* reconfigurable datacenter networks (RDCNs) such as RotorNet (SIGCOMM 2017), Opera (NSDI 2020) and Sirius (SIGCOMM 2020) have been shown to provide high throughput, by emulating a *complete graph* through fast periodic circuit switch scheduling.

However, to achieve such a high throughput, existing reconfigurable network designs pay a high price: in terms of potentially high delays, but also, as we show as a first contribution in this paper, in terms of the high buffer requirements. In particular, we show that under buffer constraints, emulating the high-throughput complete graph is infeasible at scale, and we uncover a spectrum of unvisited and attractive alternative RDCNs, which emulate regular graphs, but with lower node degree than the complete graph.

We present Mars, a periodic reconfigurable topology which emulates a d-regular graph with near-optimal throughput. In particular, we systematically analyze how the degree d can be optimized for throughput given the available buffer and delay tolerance of the datacenter. We further show empirically that Mars achieves higher throughput compared to existing systems when buffer sizes are bounded.

CCS CONCEPTS

• **Networks** → **Network architectures**; **Network performance modeling**.

KEYWORDS

Datacenter, Throughput, Buffer requirements, Reconfigurable networks.

ACM Reference Format:
Vamsi Addanki, Chen Avin, and Stefan Schmid. 2023. Mars: Near-Optimal Throughput with Shallow Buffers in Reconfigurable Datacenter Networks. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts), June 19–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3578338.3593551

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA.

ACM ISBN 979-8-4007-0074-3/23/06. https://doi.org/10.1145/3578338.3593551

1 INTRODUCTION

With the popularity of data-centric and distributed applications, the traffic in datacenters is growing explosively. Dealing with this traffic however becomes increasingly challenging: while cloud traffic roughly doubles each year, the capacity increase provided by electrical switches for a given power and cost starts to lag behind. This gap is expected to worsen with the current trend to hardware-driven workloads such as distributed machine learning training [4].

As the *throughput* of datacenter networks is becoming more and more critical for application performance, over the last years, great efforts have been made to increase the capacity of datacenter topologies. A particularly innovative architecture to meet the stringent bandwidth requirements of modern datacenters, are *reconfigurable* (optical) datacenter networks (RDCNs) [4, 6, 7], e.g., based on optical circuit switches, tunable lasers, and simple passive gratings [4]. By quickly cycling through a sequence of different topologies—typically matchings between top-of-rack (ToR) switches—RDCNs such as RotorNet [7], Opera [6] or Sirius [4] can provide periodic direct connectivity between rack pairs, at microsecond or even nanosecond granularity. A common property of these systems is that they emulate a *complete graph*, and so avoid the "bandwidth tax" of multi-hop forwarding [5, 6]. Indeed, empirical studies show that periodic reconfigurable datacenter topologies can achieve significantly higher throughput compared to cost-equivalent traditional datacenters based on *static* topologies [4, 6, 7].

This paper is motivated by the observation that the existing approach of using reconfigurable technologies to emulate complete graphs comes at a price: long delays and large buffer requirements.

First, at scale, emulating a complete graph can entail *long delays*: the denser the emulated network, the longer the periodic reconfiguration cycle and hence the longer a given rack pair has to wait to be connected again. Second, as we show analytically in this paper, the resulting long reconfiguration cycles require excessive buffering at the ToR switches and end-hosts. The required buffer can intuitively be viewed as *bandwidth-delay* product of dynamic topologies similar to the corresponding notion for static topologies in TCP literature. In practice though, datacenter switches are equipped with shallow buffers. Further, several studies in the recent past show an increasing gap between switch capacity and buffer sizes [1, 3].

Our main insight in this paper is that accounting for buffer constraints can significantly change the design considerations of reconfigurable datacenter networks.

In particular, we initiate the study of—and make the case for—reconfigurable networks which emulate graphs *of lower node degree*, uncovering an entire spectrum of possible topology designs. We present a systematic and formal analysis of the design spectrum and

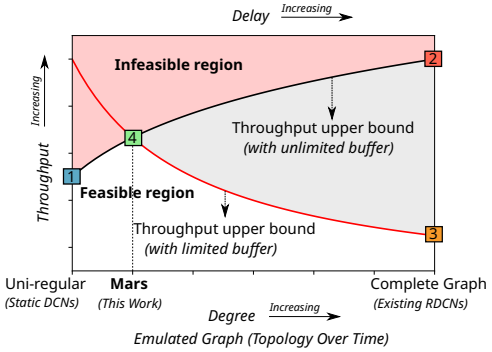

Figure 1: Periodic reconfigurable datacenter topologies pose fundamental tradeoffs across throughput, delay and buffer requirements. Existing designs are at the extremes of a spectrum of optimal designs with lower delay and buffer requirements.

the tradeoffs that these topologies introduce in terms of throughput, delay, and buffer requirements. Our perspective and the main tradeoffs are intuitively visualized in Figure 1 (left), where on the x-axis (at the bottom) we show the topologies according to the node degree of the graph they emulate. The x-axis (at the top) also represents increasing delay from left to right. The y-axis shows the throughput of the system from low to high. Let us elaborate on the important points in the figure (summarizes in the table on the right):

1 **Static DCNs:** On the very left of the design space are traditional static datacenter networks referred as uni-regular topologies e.g., expander based DCNs[1]. Such a design in principle incurs the lowest delay and requires the least amount of buffer to achieve its ideal throughput, i.e., the optimal throughput under low delay tolerance. However, since the topology remains static, for scalability reasons, each ToR switch can only connect to a limited number of other ToR switches. This results in long multi-hop paths and hence a high "bandwidth tax" [5], and lower throughput.

2 **Existing periodic RDCNs:** To reduce the "bandwidth tax" and to achieve high throughput, existing designs resort to emulating a complete graph where each rack-pair is connected directly once in every matching cycle [4, 7]. Such designs achieve high throughput, in fact maximum across all topologies in our design space, i.e., the optimal throughput. However, as we will show in this paper, the high throughput offered by emulating a complete graph comes at the cost of high delay and is subject to the availability of large buffers.

3 **Existing periodic RDCNs under resource constraints:** Given the large buffer requirements of existing designs (emulating a complete graph), we study their throughput under limited buffer. Interestingly, we find that, existing periodic reconfigurable datacenter topologies may perform equally or worse compared to a static uni-regular topology in terms of throughput, when buffer sizes are bounded.

4 **Mars:** Exploiting the fundamental tradeoffs across the design space, we propose Mars, a periodic reconfigurable topology that provides near optimal, high throughput with the limited amount of available buffer. Specifically, we parametrize our design based on the delay tolerance and available buffer. We systematically determine the optimal degree d which depending on the resource constraints lies between a static topology and a complete graph.

It is interesting to observe from Figure 1 that the throughput-delay relation implies an *infeasible* region for topology design (shown in red shade). The available buffer space further restricts the design space (shown in gray shade), imposing a fundamental tradeoff across the topologies within the feasible region in terms of throughput, delay and buffer.

Our analytical approach in this paper is novel and relies on a reduction of a periodic evolving graph to a specific static graph. This enables us to study the throughput maximization problem using well-known graph analysis techniques for static graphs and to analytically evaluate the throughput of both existing dynamic topologies (RotorNet, Opera, and Sirius) as well as possible alternatives (Mars). We believe that this reduction technique may be of independent interest and could be potentially used to study other properties of dynamic topologies.

The full version of this paper appears in [2].

ACKNOWLEDGEMENTS

We would like to thank our shepherd, Mohammad Hajiesmaili, as well as the anonymous reviewers and the technical program committee of SIGMETRICS 2023 for their useful feedback in shaping the final version of this paper. This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, consolidator project Self-Adjusting Networks (AdjustNet), grant agreement No. 864228, Horizon 2020, 2020-2025.

REFERENCES

[1] Vamsi Addanki, Maria Apostolaki, Manya Ghobadi, Stefan Schmid, and Laurent Vanbever. Abm: Active buffer management in datacenters. In *Proceedings of the ACM SIGCOMM 2022 Conference*, 2022.
[2] Vamsi Addanki, Chen Avin, and Stefan Schmid. Mars: Near-optimal throughput with shallow buffers in reconfigurable datacenter networks. *Proc. ACM Meas. Anal. Comput. Syst.*, 7(1), mar 2023.
[3] Wei Bai, Shuihai Hu, Kai Chen, Kun Tan, and Yongqiang Xiong. One more config is enough: Saving (dc)tcp for high-speed extremely shallow-buffered datacenters. *IEEE/ACM Transactions on Networking*, 29(2):489–502, 2021.
[4] Hitesh Ballani, Paolo Costa, Raphael Behrendt, Daniel Cletheroe, Istvan Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen, and Hugh Williams. Sirius: A flat datacenter network with nanosecond optical switching. In *Proceedings of the ACM SIGCOMM 2020 Conference*, page 782–797, 2020.
[5] Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin. Cerberus: The power of choices in datacenter topology design - a throughput perspective. *Proc. ACM Meas. Anal. Comput. Syst.*, 5(3), dec 2021.
[6] William M. Mellette, Rajdeep Das, Yibo Guo, Rob McGuinness, Alex C. Snoeren, and George Porter. Expanding across time to deliver bandwidth efficiency and low latency. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 1–18, Santa Clara, CA, February 2020. USENIX Association.
[7] William M. Mellette, Rob McGuinness, Arjun Roy, Alex Forencich, George Papen, Alex C. Snoeren, and George Porter. Rotornet: A scalable, low-complexity, optical datacenter network. In *Proceedings of the ACM SIGCOMM 2017 Conference*, page 267–280, 2017.

[1] These works also claimed that static expander-based topologies are similar or better designs (in terms of performance and cost) compared to Clos-based topologies. We henceforth focus on static expanders. A detailed discussion can be found in the full version of this paper [2].